Empirical Studies in Discourse

Empirical Studies in Discourse

Marilyn A. Walker* & Johanna D. Moore †

1. Introduction

Computational theories of discourse are concerned with the context-based interpretation or generation of discourse phenomena in text and dialogue. In the past, research in this area focused on specifying the mechanisms underlying particular discourse phenomena; the models proposed were often motivated by a few constructed examples. While this approach led to many theoretical advances, models developed in this manner are difficult to evaluate because it is hard to tell whether they generalize beyond the particular examples used to motivate them.

Recently however the field has turned to issues of robustness and the coverage of theories of particular phenomena with respect to specific types of data. This new empirical focus is supported by several recent advances: an increasing theoretical consensus on discourse models; a large amount of online dialogue and textual corpora available; and improvements in component technologies and tools for building and testing discourse and dialogue testbeds. This means that it is now possible to determine how representative particular discourse phenomena are, how frequently they occur, whether they are related to other phenomena, what percentage of the cases a particular model covers, the inherent difficulty of the problem, and how well an algorithm for processing or generating the phenomena should perform to be considered a good model.

This issue brings together a collection of papers illustrating recent approaches to empirical research in discourse generation and interpretation. Section 2 gives a general overview of empirical studies in discourse and describes an empirical research strategy that leads from empirical findings to general theories. Section 3 discusses how each article exemplifies the empirical research strategy and how empirical methods have been employed in each research project.

2. Why Empirical Studies?

What is the role of empirical studies in research on computational models of discourse? We believe that developing *general* theories depends on the ability to characterize computational models of discourse, and their behaviors, tasks, and tasks contexts, in terms of sets of *features*, that can be used to make and evaluate predictions about what affects the behavior under investigation (Cohen, 1995; Sparck-Jones and Galliers, 1996; Walker, 1996).[1] The role of empirical methods is to help researchers discover general features by analyzing specific discourse phenomena or programs that interpret or generate them.

* AT&T Labs Research, 600 Mountain Ave. 2D-441, Murray Hill, NJ. 07974, walker@research.att.com
† Computer Science Dept.,University of Pittsburgh, jmoore@cs.pitt.edu

1 Sparck-Jones and Galliers (1996, p.23) call features *performance factors* and distinguish between *environmental factors* which are features of the task that are fixed from the system designer's viewpoint, *system factors* which reflect design choices, algorithm features, or other input factors, and *system effects* which are features that characterize the behavior of the system.

Once relevant features are identified, hypotheses about the relationship between them can be formed, and controlled studies that test the hypothesized relationships can be devised. This approach leads to general theories via the following steps, which many readers will recognize as a variation of Cohen's EMPIRICAL GENERALIZATION STRATEGY (Cohen, 1995, p.6):

1. **Feature identification:** identify features of the discourse, tasks, and context that may influence the target behavior;
2. **Modeling:** develop a causal model of how these features influence the target behavior;
3. **Evaluation:** assess the performance of the model (often implemented in a program) for producing the target behavior on the tasks and in the contexts for which it was devised;
4. **Generalization:** once the model makes accurate predictions, generalize the features so that other discourses, tasks, and contexts are encompassed by the causal model, and test whether the general model accurately predicts behavior in the larger set of discourses, tasks, and contexts.

This strategy provides a general research methodology in which the study of discourse phenomena proceeds in several stages, each of which employ empirical methods. The first stage is to identify features that may affect the behavior of interest. Hypotheses about what features are relevant may come from analysis of a corpus of naturally occurring discourses, a study of the literature on the phenomena of interest, an analysis of tasks in a particular domain or the analysis of a program that exhibits a behavior of interest. These hypotheses are then used to develop or refine a causal model of how the features influence the target behavior. At this point, the model can be evaluated in many ways, e.g. a program that embodies the model can be implemented and evaluated. If evaluation indicates that performance of the model is not satisfactory, further work must be done at stages 1 and 2 to identify features that influence the target behavior and to model their interactions appropriately.

Stages 3 and 4 are key to the ability to generalize; once researchers are able to derive quantitative results by testing a model against a data set, general theories can arise from qualitatively analyzing which aspects of the model most directly affect the desired behavior, and from evaluating the model in qualitatively different situations. For this reason, much recent work has been concerned with methodological issues of how to quantitatively measure performance.

But how do empirical methods help researchers discover general features and generate hypotheses? Recent work uses some combination of the following empirical methods: (1) Tagging of discourse phenomena in corpora; (2) Induction of algorithms or discourse models from tagged data; (3) Comparison of algorithm output to human performance; (4) Human scoring of an algorithm's output; (5) Task efficacy evaluation based on the the domain; (6) Ablation studies where algorithm features are systematically turned off. (7) Wizard of Oz studies; (8) Testbeds of (parameterizable) dialogue models using computer-computer dialogue simulation; (9) Testbeds of (parameterizable) dialogue models implemented in human-computer dialogue interfaces. How are these methods used and how do they contribute to the development of general theories?

Discourse tagging classifies discourse units in naturally occurring texts or dialogues into one of a set of categories. Discourse units range from referring expressions and syntactic constructions (Fox, 1987; Kroch and Hindle, 1982; Prince, 1985), to words or

phrases (Heeman and Allen, 1994; Hirschberg and Litman, 1993; Novick and Sutton, 1994), to utterances and relationships among them (Dahlback, 1991; Reithinger and Maier, 1995; Moser and Moore, 1995; Nagata, 1992; Rose et al., 1995), to multi-utterance units identified by a range of criteria such as speaker intention or initiative (Flammia and Zue, 1995a; Hirschberg and Nakatani, 1996; Whittaker and Stenton, 1988). The article by Carletta *et. al.*(this volume) presents a tagging scheme for three levels of discourse structure.

Discourse tags categorize **either** features of the input (independent variables) or features of the output (dependent variables). Often hypotheses about input features that affect the target behavior are found in previous work (stage 1 of the methodology). In this case, the tagging contributes to developing a causal model. The tagged corpus is used to test whether the features predict the target behavior. For example, researchers have devised algorithms for generating the surface form of explanations and instructions from underlying intention-based representations by tagging naturally occurring discourses for surface form features, informational relations, and intentional relations (Vander Linden and Di Eugenio, 1996; Moser and Moore, 1995; Moore and Pollack, 1992; Paris and Scott, 1994). Another promising area is speech act (dialogue move) tagging, where, for example, researchers have tested whether an automatic tagger trained on the tagged corpus can improve the performance of a speech recognizer with tag-specific language models (Taylor et al., 1996), and whether an induced discourse model based on the tagging can predict the **next** dialogue act in the dialogue, and thus affect how the system translates the next utterance (Reithinger and Maier, 1995; Nagata, 1992).

Tagging is also critical for ablation studies, where algorithm features are selectively turned off, and performance examined. Tagging can characterize the input in terms of features the algorithm uses for producing the target behavior or characterize the target behavior. For example, Lappin and Leass (1994) report an ablation study of an anaphora resolution algorithm, operating on computer manuals, in which various factors that were hypothesized to determine the most likely antecedent were selectively turned off. (See also Dagan et al. (1995)). Sample results include the finding that there is no effect on performance, for this type of texts, when an antecedent's likelihood is increased for parallelism.

Another use of discourse tagging is for algorithm induction using automatic classifiers, such as C4.5 or CART, that produce decision trees from data sets described by a set of features (Brieman et al., 1984; Quinlan, 1993). This approach uses automatic methods for stages 2 and 3 of our empirical research strategy. Since discourse tagging associates sets of features with discourse phenomena, tagged data is used as input to these automatic classifiers. The decision tree produced by the classifiers functions as a causal model, which can then be examined and further tested. For example, this method has been used to identify features for predicting accent assignment in text-to-speech (Hirschberg, 1993), for repairing disfluencies (Nakatani and Hirschberg, 1993), cue vs. noncue uses of discourse cue words (Litman, 1996; Siegel and McKeown, 1994), discourse segment boundaries (Grosz and Hirschberg, 1992),(Passonneau and Litman, this volume), intonational phrase boundaries (Wang and Hirschberg, 1992), and the most likely antecedents for anaphors (Aone and Bennet, 1995; Connolly, Burger, and Day, 1994).

Discourse tagging is also instrumental in stage 3 of our empirical method, by producing a test set that can be used for comparison to a program's output. This method is used by most of the articles in this volume. A common application is testing coreference resolution algorithms; the tags indicate the preferred interpretation of a potentially ambiguous utterance containing anaphoric noun phrases (Walker, 1989; Suri and McCoy, 1994; Chinchor and Sundheim, 1995). Coreference algorithms are then tested on their

ability to select the right equivalence class for an anaphoric noun phrase. The same method has been applied to empirically testing an algorithm for resolving verb phrase ellipsis (Hardt, 1992).

In each case, we can generalize on the basis of specific features from studies of specific algorithms operating on specific corpora, whenever the corpora represent a general task for the algorithm. The more varied the test data is, the more generalizable we expect the results to be. For example, the claim that a model is general can be supported by test corpora representing different genres (Fox, 1987; Kroch and Hindle, 1982), or different language families, as in (Strube and Hahn, 1996; Iida, 1997; Di Eugenio, 1997; Hoffman, 1997).

Human performance can also be compared to algorithm output through the use of reaction time or comprehension or production experiments (Brennan, 1995; Gordon, Grosz, and Gilliom, 1993; Hudson-D'Zmura, 1988). These methods allow researchers fine-grained control of the phenomena studied, and avoid problems with sparse data that can arise with corpus analyses. Reaction time studies also provide researchers with an indirect measure of how humans process a particular phenomenon; processing times can then be compared with the predictions of a model.

The method of having humans score the result of an algorithm, or compare it to human performance, is useful when the algorithm output is hard to classify as an element of a finite set, e.g. when the output is a summary or an explanation produced by a natural language generation system (McKeown and Radev, 1995; Robin and McKeown, 1996; Sparck-Jones, 1993). This is the method used by Lester and Porter (this volume).

The method of task efficacy evaluation tests the effectiveness of the target behavior in an environment in which it is embedded (Sparck-Jones and Galliers, 1996). For example, in evaluating the generation mechanism of a tutorial system, a task efficacy measure evaluates how well students comprehend or learn an explanation or an instruction. If the execution of an instruction can be monitored, performance metrics can be collected (Biermann and Long, 1996; Young, 1997). Another approach is based on field studies with actual end-users (Jarke et al., 1985). Like human scoring, this method appears to be well-suited for research with complex outputs such as instructional dialogue systems.

Because testing dialogue systems requires a fully implemented natural language system, there are two empirical methods for testing hypotheses about discourse models that are independent of the current state of the art in speech or language processing. The first method is Wizard-of-Oz simulation, and the second is computational testbeds for dialogue simulation.

In the Wizard-of-Oz (WOZ) approach, a human wizard simulates the behavior of a program interacting with a human to carry out a particular task in a particular domain (Dahlback and Jonsson, 1989; Hirschman et al., 1993; Oviatt and Cohen, 1989; Whittaker and Stenton, 1989). The WOZ method can, in principle, be used to test, refine or generalize any behavior implementable in a program and thus is appropriate at several stages of our methodology. For example, the wizard may follow a protocol that includes particular system limitations or error handling strategies, to explore potential problems before implementation, e.g. determining how the program's level of interactivity affects the complexity of the instructions it is given (Oviatt and Cohen, 1989). In addition, WOZ is often used to collect sample dialogues that are not affected by system limitations; the human wizard can simulate behavior that would result in a system error, so that the resulting corpus of dialogues is not affected by human adaptation to the system's limitations. In some cases, the resulting corpus provides training data for spoken language systems (Hirschman et al., 1993), is used as a target for improved systems (Moore, Lemaire, and Rosenblum, 1996), or forms a test set for evaluating the performance of an existing natural language system (Whittaker and Stenton, 1989; Hirschman et al., 1993).

Dialogue simulation testbeds evaluate specific models of agent communicative action. This approach spans stages 1-3 of the methodology: human-human dialogues are used to formulate hypotheses about features that determine a target behavior; a program is designed in which the target behavior is parameterizable and in which metrics for evaluating performance can be collected; and then simulations are run to determine which behaviors, determined by the parameter settings, affect performance on the task. Hanks, Pollack, and Cohen (1993) discuss the role of this method in *exploration*, *confirmation* and *generalization* of particular models of agent behavior. In discourse studies, these testbeds have been used to investigate the utility of planning as an underlying model of dialogue (Power, 1979; Houghton and Isard, 1985), the interaction of risk-taking dialogue strategies and corresponding repair strategies (Carletta, 1992), the relationship between resource bounds, task complexity and dialogue strategies (Walker, 1996; Jordan and Walker, 1996), the role of belief revision in tasks in which agents negotiate a problem solution (Logan, Reece, and Sparck Jones, 1994), and the relationship between mixed initiative and knowledge distribution (Guinn, 1994).

Empirical research in testbeds for human-computer dialogue interfaces is very recent (Hirschman et al., 1990; Allen et al., 1996). This method supports studies of the interaction of various components; for example Danieli and Gerbino (1995) propose an *implicit recovery* metric for evaluating how the dialogue manager overcomes limitations in the speech recognizer. Other research parameterizes the dialogue manager to select different behaviors in different contexts, such as *expert* vs. *novice* discourse strategies (Kamm, 1995), different repair strategies (Hirschman and Pao, 1993), or different degrees of initiative (Potjer et al., 1996), (Smith and Gordon, this volume). These dialogue interfaces also provide an opportunity for task efficacy evaluation, as discussed above, and there are already many examples of dialogue systems being tested in field trials with representative populations of users (Danieli and Gerbino, 1995; Kamm, 1995; Meng et al., 1996; Sadek et al., 1996) *inter alia*.

3. Overview of the Issue

The articles in this issue represent many of the empirical methods discussed above and span research on dialogue tagging, generation of referring expressions, generation of explanations, topic identification, identification of changes in speaker intention, and the effect of initiative on the structure of dialogues. Below we discuss each article in turn, both in terms of the methods it uses and in terms of how it contributes to the development of general theories.

3.1 Carletta, Isard, Isard, Kowtko, Doherty-Sneddon and Anderson

Discourse tagging is a key component of much empirical work in discourse, however the development of discourse tag sets is still a relatively new area of endeavor. Carletta *et. al.* discuss a scheme for tagging human-human dialogues with three tag sets: *transactions*, conversational *games*, and dialogue *moves*. A *transaction* is a subdialogue that accomplishes one major step in the participants' plan for achieving the task (Isard and Carletta, 1995). The size and shape of transactions is therefore largely dependent on the task. Each transaction consists of a sequence of conversational *games*, where a conversational game is a set of utterances starting with an initiation (e.g., a request for information) and encompassing all successive utterances until the purpose of the game has been fullfilled or the game has been abandoned. Each game consists of a sequence of *moves*, each of which is classified as either an initiative (e.g., instruction, explanation) or a response (e.g., reply, acknowledgment).

The paper discusses issues with determining the reliability and generality of tagging

sets, and with refining tagsets on the basis of reliability data (Carletta, 1996; Condon and Cech, 1995). The Discourse Working Group is also applying this tagging scheme to other dialogue types to evaluate its generality (Hirschman et al., 1996; Luperfoy, 1996).

3.2 Hearst

The target behavior that Hearst is concerned with is subtopic identification in expository texts. She tests the hypothesis that TERM REPETITION is a primary feature of textual cohesion (Morris and Hirst, 1991) by using it as the basis for two different algorithms for identifying multiparagraph subtopical units. The algorithms are evaluated by comparing the units they propose against a baseline of randomly generated topic boundaries, and against a corpus tagged by human judges. Precision, recall, and κ (Carletta, 1996; Krippendorf, 1980) are used as evaluation metrics to assess the performance of the two algorithms.

In order to generalize, Hearst tests the algorithm's performance on a new task: that of distinguishing boundaries between sets of concatenated news articles. Hearst's algorithm performs comparably to other algorithms on the new task, showing that term repetition may be a more general indicator of subtopic boundaries. Future work could test further generalizations; term repetition may also indicate subtopics in other discourse or dialogue environments, and may interact with other features that correlate with topic boundaries, such as pauses, intonation or cue words (Cahn, 1992; Hirschberg and Nakatani, 1996).

3.3 Lester and Porter

The target behavior that concerns Lester and Porter is generating paragraph-length explanations in the biology domain. Given a goal to explain a biology concept or process, their KNIGHT system selects relevant information from a large knowledge base, organizes it, and then generates it. It is clearly not possible to evaluate how well KNIGHT produces coherent paragraph-length explanations by comparing KNIGHT's explanations word for word, or even proposition for proposition, with a corpus of human explanations. There are simply too many choices on the path from knowledge base to surface form. So Lester and Porter compare KNIGHT to human performance by having domain experts score a corpus consisting of both KNIGHT and human explanations. The domain experts are unaware of the fact that some explanations are generated by a computer. KNIGHT's performance is evaluated on the basis of grading explanations on a scale of A to F for the features of coherence, content, organization, writing style and correctness.

To generalize their results, Lester and Porter propose further fine-grained analyses of KNIGHT's output by sentential form or referring expressions. Other generalizations could arise by showing that KNIGHT's content organization operators (EDPs) could be used in other domains for generating explanations, as in (Robin and McKeown, 1996).

3.4 Passonneau and Litman

The target behavior that Passonneau and Litman's article models is the identification of multi-utterance units in spoken story narratives that correspond to speaker's intention. In order to define a test set for the target behavior, 7 human subjects tagged each utterance in a narrative as a *boundary* where the speaker starts communicating a new intention, or as a *non-boundary*, where the speaker continues discussing the current intention. The boundaries marked by either 3 or 4 subjects (out of 7) are used to define the target behavior for boundary identification. Passonneau and Litman then examine three classes of utterance features to determine correlations with boundary and non-boundary utterances: (1) coreferential and inferential relationships between noun phrases across two adjacent utterances; (2) the occurrence of discourse cue phrases at the beginning of an

utterance; and (3) prosodic/acoustic utterance features such as phrase-final intonation and utterance-initial pauses. They develop and evaluate three algorithms for producing the target behavior from these features, two hand-developed and one automatically induced.

Generalizations of this work arise from other research on different speech genres in different environments that also found that coreferential relationships, pausing, and intonation are correlated with discourse structure (Cahn, 1992; Fox, 1987; Hirschberg and Nakatani, 1996). Future work can further test the generalizability of the results reported here: the features used could be examined in other types of spoken monologues, in texts, and in dialogue.

3.5 Smith and Gordon

Smith and Gordon examine the effect of initiative on dialogue structure in dialogues in which human subjects interact with the computer to diagnose and repair problems with simple circuits. Initiative is varied by setting the system's initiative to DIRECTIVE or DECLARATIVE mode. In directive mode, the system instructs the student at each step. In declarative mode, the system lets the student take the initiative, but volunteers relevant facts. Dialogue structure is tagged via a model that segments circuit-repair dialogues into five phases: *introduction, assessment, diagnosis, repair* and *test*. Then Smith and Gordon examine how the subdialogue length varies depending on initiative mode.

Their results exemplify the empirical generalization strategy by showing that a subdialogue model based on WOZ simulations can be generalized to human-computer dialogues. They also show that claims about the effect of initiative on dialogue structure in human-human dialogues in other domains (Whittaker and Stenton, 1988; Walker and Whittaker, 1990) generalize to human-computer dialogues in the circuit repair domain. Further generalizations could result from determining whether the subdialogue model can be used in other types of human-human or human-computer problem-solving dialogues.

3.6 Yeh and Mellish

The target behavior that Yeh and Mellish model is the generation of anaphoric noun phrases in Chinese texts. The algorithm must select from among zero pronouns, overt pronouns and full noun phrases; in addition, for full noun phrases, appropriate content must be determined. Their training set is a corpus of Chinese texts tagged for anaphoric noun phrases and for features claimed to affect noun-phrase form. They construct a decision tree by sequentially allowing additional features to affect decisions on anaphoric form, wherever there is room for improvement.

In order to test the derived decision tree, they construct a test set of texts generated by a Chinese generator. At each location where a noun phrase occurs, a set of choices of nominal referring expressions are given. Then they conduct two comparisons. First, human judges select from among the forms and the selections of human judges are compared among themselves. Second, a program implementing the derived decision tree selects among the forms and the program's behavior is compared to the human judges.

This work could be generalized by comparing their predictive features with those used in algorithms for generating referring expressions in English (Passonneau, 1995). Other generalizations could arise from comparing the decision trees with factors that affect anaphoric forms in Japanese (Iida, 1997), Italian (Di Eugenio, 1997), or Turkish (Hoffman, 1997). In addition, it would be useful to test their suggestion that the lack of agreement among native speakers as to the preferred form of anaphoric noun phrases was partially determined by the use of texts generated by a natural language generator.

It is possible that the same experiment carried out on naturally occurring texts would generate a similar amount of disagreement.

4. Future Directions

In recent years, there has clearly been a groundswell of interest in empirical methods for analyzing discourse. A survey of recent ACL papers shows that the percentage of empirical papers in semantics, pragmatics and discourse hovered between 8% and 20% until 1993 when it increased to 40%. In 1995 and 1996, 75% of the ACL papers in semantics, pragmatics and discourse used empirical methods. While a great deal of progress has been made, several obstacles impede empirical research. The discourse community must develop more shared methods, tools and resources.

First, researchers in discourse must agree on methods for quantitatively characterizing performance and on ways to determine whether the metrics are serving their intended diagnostic function (Moore and Walker, 1995; Cohen, 1995; Sparck-Jones and Galliers, 1996). Recent work includes discussion of appropriate statistical methods and metrics for spoken dialogue systems (Bates and Ayuso, 1993; Danieli et al., 1992; Hirschman et al., 1990; Hirschman and Pao, 1993; Simpson and Fraser, 1993), information extraction systems (Lewis, 1991; Chinchor, Hirschman, and Lewis, 1993; Chinchor and Sundheim, 1995), and tagging reliability (Carletta, 1996).

Second, we must develop more shared tools. The lack of tools greatly increases the cost of accurate coding, which could be reduced with coding tools that structure the coder's input and checks that it it is within the coding scheme's constraints. To date most coders enter data by hand in a word processor or using home-grown, hastily constructed tools. To our knowledge, there is only one publicly available tool for dialogue structure coding (Flammia and Zue, 1995b).

Third, we must increase the number of and representativeness of dialogue and text corpora. To our knowledge, there are no publicly available human-computer dialogue corpora, nor are there human-human dialogues representing a broad range of spoken-dialogue applications. Similarly, there are no publicly available corpora of text-based explanations in particular domains that could be a resource for the generation community. However, even if more corpora become available, most discourse studies require data to be tagged, and there are currently no publicly available tagged corpora. In order to develop a large shared resource of tagged materials, the discourse community must share efforts across sites. We need to develop shared coding schemes and make coded data publicly available to support comparisons of different models. The community is currently addressing these issues through a series of working groups (Hirschman et al., 1996; Luperfoy, 1996). Given the current state of the art, we expect these issues to concern the community for some time.

5. Acknowledgments

We are greatly indebted to the many people who contributed to this special issue by serving as reviewers for the 29 papers that were submitted. We would also like to thank Lynette Hirschman, Aravind Joshi and Marti Hearst for helping us organize the AAAI Workshop on Empirical Methods in Discourse that provided the impetus for this issue.

References

Allen, James F., Bradford W. Miller, Eric K. Ringger, and Teresa Sikorski. 1996. A robust system for natural spoken dialogue. In *Association of Computational Linguistics Annual Meeting*, pages 62–70.

Aone, Chinatsu and Scott Bennet. 1995. Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, pages 122–129.

Bates, Madeleine and Damaris Ayuso. 1993. A proposal for incremental dialogue evaluation. In *Proceedings of the Darpa Speech and NL Workshop*, pages 319–322.

Biermann, A. W. and Philip M. Long. 1996. The composition of messages in speech-graphics interactive systems. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, pages 97–100.

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive processes*, 10(2):137–167.

Brieman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey California.

Cahn, Janet. 1992. An investigation into the correlation of cue phrases, unfilled pauses and the structuring of spoken discourse. In *Workshop on Prosody in Natural Speech*, pages 19–31. Institute for Research in Cognitive Science, University of Pennsylvania, TR IRCS-92-37.

Carletta, Jean C. 1992. *Risk Taking and Recovery in Task-Oriented Dialogue*. Ph.D. thesis, Edinburgh University.

Carletta, Jean C. 1996. Assessing the reliability of subjective codings. *Computational Linguistics*, 20(4).

Chinchor, Nancy, Lynette Hirschman, and David D. Lewis. 1993. Evaluating message understanding systems: An analysis of the third message understanding conference (muc-3). *Computational Linguistics*, 19(3):409–451.

Chinchor, Nancy and Beth Sundheim. 1995. Message understanding conference MUC tests of discourse processing. In *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 21–27.

Cohen, Paul. R. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press, Boston.

Condon, Sherri L. and Claude G. Cech. 1995. Functional comparison of face-to-face and computer-mediated decision-making interactions. In Susan Herring, editor, *Computer-Mediated Converstaion*. John Benjamins.

Connolly, Dennis, John D. Burger, and David S. Day. 1994. A machine learning approach to anaphoric reference. In *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP)*.

Dagan, Ido, John Justeson, Shalom Lappin, Herbert Leass, and Amnon Ribak. 1995. Syntax and lexical statistics in anaphora resolution. *Applied Artificial Intelligence - An International Journal*, 9(4):633–644.

Dahlback, Nils. 1991. *Representations of Discourse: Cognitive and Computational Aspects*. Ph.D. thesis, Linkoping University.

Dahlback, Nils and Arne Jonsson. 1989. Empirical studies of discourse representations for natural language interfaces. In *Proc. 4th Conference of the European Chapter of the ACL, Association of Computational Linguistics*, pages 291–298.

Danieli, M., W. Eckert, N. Fraser, N. Gilbert, M. Guyomard, P. Heisterkamp, M. Kharoune, J. Magadur, S. McGlashan, D. Sadek, J. Siroux, and N. Youd. 1992. Dialogue manager design evaluation. Technical Report Project Esprit 2218 SUNDIAL, WP6000-D3.

Danieli, Morena and Elisabetta Gerbino. 1995. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39.

Di Eugenio, Barbara. 1997. Centering theory and the italian pronominal system. In *Centering in Discourse*. Oxford University Press.

Flammia, Giovanni and Victor Zue. 1995a. Empirical results of dialogue coding. In *EUROSPEECH 95*.

Flammia, Giovanni and Victor Zue. 1995b. N.b.: A graphical user interface for annotating spoken dialogue. In Marilyn Walker and Johanna Moore, editors, *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*.

Fox, Barbara A. 1987. *Discourse Structure and Anaphora: Written and Conversational English*. Cambridge University Press.

Gordon, Peter C., Barbara J. Grosz, and Laura A. Gilliom. 1993. Pronouns, names and the centering of attention in discourse. *Cognitive Science*, 17(3):311–348.

Grosz, Barbara J. and Julia B. Hirschberg. 1992. Some intonational characteristics of discourse structure. In *ICSLP*.

Guinn, Curry I. 1994. *Meta-Dialogue Behaviors: Improving the efficiency of Human-Machine Dialogue*. Ph.D. thesis, Duke University.

Hanks, Steve, Martha Pollack, and Paul Cohen. 1993. Benchmarks, testbeds, controlled experimentation and the design of agent architectures. *AI Magazine*, December.

Hardt, Daniel. 1992. An algorithm for vp ellipsis. In *Annual Meeting of the Association of Computational Linguistics*, pages 9–14.

Heeman, Peter A. and James Allen. 1994. Detecting and correcting speech repairs. In *ACL94*.

Hirschberg, Julia and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530.

Hirschberg, Julia and Christine Nakatani. 1996. A prosodic analysis of discourse segments in direction-giving monologues. In *Annual Meeting of the Association of Computational Linguistics*, pages 286–293.

Hirschberg, Julia B. 1993. Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence Journal*, 63:305–340.

Hirschman, L., M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann. 1993. Multi-site data collection and evaluation in spoken language understanding. In *Proceedings of the Human Language Technology Workshop*, pages 19–24.

Hirschman, Lynette, Deborah A. Dahl, Donald P. McKay, Lewis M. Norton, and Marcia C. Linebarger. 1990. Beyond class a: A proposal for automatic evaluation of discourse. In *Proceedings of the Speech and Natural Language Workshop*, pages 109–113.

Hirschman, Lynette, Aravind Joshi, Johanna Moore, and Marilyn Walker. 1996. IRCS Workshop on Discourse Tagging. Technical report, University of Pennsylvania, `http://www.cis.upenn.edu:80/~ircs/discourse-tagging/toplevel.html`.

Hirschman, Lynette and Christine Pao. 1993. The cost of errors in a spoken language system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pages 1419–1422.

Hoffman, Beryl. 1997. Word order, information structure and centering in turkish. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse*. Oxford University Press.

Houghton, G. and S. Isard. 1985. Why to speak, what to say and how to say it : Modelling language production in discourse. In *Proceedings of the International Workshop on Modelling Cognition*, University of Lancaster.

Hudson-D'Zmura, Susan B. 1988. *The Structure of Discourse and Anaphor Resolution: The discourse Center and the Roles of Nouns and Pronouns*. Ph.D. thesis, University of Rochester.

Iida, Masayo. 1997. Discourse coherence and shifting centers in japanese texts. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince, editors, *Centering in Discourse*. Oxford University Press.

Isard, Amy and Jean C. Carletta. 1995. Replicability of transaction and action coding in the map task corpus. In Marilyn Walker and Johanna Moore, editors, *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*.

Jarke, Matthias, Jon A. Turner, Edward A. Stohr, Yannis Vassiliou, Norman H. White, and Ken Michielsen. 1985. A field evaluation of natural language for data retrieval. *IEEE Transactions on Software Engineering*, SE-11, No.1:97–113.

Jordan, Pamela and Marilyn A. Walker. 1996. Deciding to remind during collaborative problem solving: Empirical evidence for agent strategies. In *Conference of the American Association of Artificial Intelligence, AAAI*.

Kamm, Candace. 1995. User interfaces for voice applications. In David Roe and Jay Wilpon, editors, *Voice Communication between Humans and Machines*. National Academy Press, pages 422–442.

Krippendorf, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, Ca.

Kroch, Anthony S. and Donald M. Hindle. 1982. A quantitative study of the syntax of speech and writing. Technical report, University of Pennsylvania.

Lappin, Shalom and Herbert Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–562.

Lewis, David D. 1991. Evaluating text categorization. In *DARPA Speech and Natural Language Workshop*, pages 312–318.

Litman, Diane. 1996. Cue phrase classification using machine learning. *Journal of Artificial Intelligence Research*, 5:53–94.

Logan, Brian, Steven Reece, and Karen Sparck Jones. 1994. Modelling information retrieval agents with belief revision. In *Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 91–100, London. Springer-Verlag.

Luperfoy, Susan. 1996. Discourse Resource Initiative Home Page. Technical report, `http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html`.

McKeown, Kathleen R. and Dragomir R. Radev. 1995. Generating summaries of multiple news articles. In Edward A. Fox, Peter Ingwersen, and Raya Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, July.

Meng, Helen, Senis Busayapongchai, James Glass, Dave Goddeau, lee Hetherington, Ed Hurley, Christine Pao, Joe Polifroni, Stephanie Seneff, and Victor Zue. 1996. Wheels: A conversational system in the automobile classifieds domain. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, pages 165–168.

Moore, Johanna and Marilyn Walker. 1995. Proceedings of the 1995 aaai workshop on empirical methods in discourse interpretation and generation. Technical report, AAAI: `http://www.aaai.org/Publications/TechReports/reportcatalog.html#spring`.

Moore, Johanna D., Benoît Lemaire, and James A. Rosenblum. 1996. Discourse generation for instructional applications: Identifying and exploiting relevant prior explanations. *Journal of the Learning Sciences*, 5(1):49–94.

Moore, Johanna D. and Martha E. Pollack. 1992. A problem for rst: The need for multi-level discourse analysis. *Computational Linguistics*, 18(4).

Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Moser, Margaret G. and Johanna Moore. 1995. Investigating cue selection and placement in tutorial discourse. In *ACL 95*.

Nagata, Masaaki. 1992. Using pragmatics to rule out recognition errors in cooperative task-oriented dialogues. In *ICSLP*.

Nakatani, Christine H. and Julia Hirschberg. 1993. A speech-first model for repair detection and correction. In *Association of Computational Linguistics Annual Meeting*, pages 46–53.

Novick, David and Stephen Sutton. 1994. An empirical model of acknowledgment for spoken language systems. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*.

Oviatt, Sharon L. and Philip R. Cohen. 1989. The effects of interaction on spoken discourse. In *Proc. 27th Annual Meeting of the Association of Computational Linguistics*, pages 126–134.

Paris, Cecile and Donia Scott. 1994. Stylistic variation in multilingual instructions. In *The 7th International Conference on Natural Language Generation*.

Passonneau, Rebecca J. 1995. Integrating gricean and attentional constraints. In *IJCAI 95*.

Potjer, J., A. Russel, L. Boves, and E. den Os. 1996. Subjective and objective evaluation of two types of dialogues in a call assistance service. In *1996 IEEE Third Workshop: Interactive Voice Technology for Telecommunications Applications, IVTTA*. IEEE, pages 89–92.

Power, R. 1979. The organisation of purposeful dialogues. *Linguistics*, 17:107–152.

Prince, Ellen F. 1985. Fancy syntax and shared knowledge. *Journal of Pragmatics*, pages 65–81.

Quinlan, J. Ross. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Reithinger, Norbert and Elisabeth Maier. 1995. Utilizing statistical speech act processing in verbmobil. In *ACL 95*.

Robin, J. and K. McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, 85. Special Issue on Empirical Methods.

Rose, Carolyn Penstein, Barbara Di Eugenio, Lori Levin, and Carolyn Van Ess-Dykema. 1995. Discourse processing of dialogues with multiple threads. In *ACL*.

Sadek, M. D., A. Ferrieux, A. Cosannet, P. Bretier, F. Panaget, and J. Simonin. 1996. Effective human-computer cooperative spoken dialogue: The ags demonstrator. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, pages 169–173.

Siegel, Eric V. and Kathleen R. McKeown. 1994. Emergent linguistic rules from inducing decisions trees: Disambiguating discourse clue words. In *AAAI94*, pages 820–826.

Simpson, A. and N. A. Fraser. 1993. Black box and glass box evaluation of the sundial system. In *Proceedings of the Third European Conference on Speech Communication and Technology*, pages 1423–1426.

Sparck-Jones, Karen. 1993. What might be in a summary? In *Proceedings of Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Universitatsverlag Knstanz.

Sparck-Jones, Karen and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems*. Springer.

Strube, Michael and Udo Hahn. 1996. Functional centering. In *Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics, ACL*.

Suri, Linda Z. and Kathleen F. McCoy. 1994. RAFT/RAPR and Centering: A comparison and discussion of problems related to processing complex sentences. *Computational Linguistics*, 20(2).

Taylor, Paul, Hiroshi Shimodaira, Stephen Isard, Simon King, and Jaqueline Kowtko. 1996. Using prosodic information to constrain language models for spoken dialogue. In *ICSLP96*.

Vander Linden, Keith and Barbara Di Eugenio. 1996. A corpus study of negative imperatives in natural language instructions. In *COLING96*.

Walker, Marilyn A. 1989. Evaluating discourse processing algorithms. In *Proc. 27th Annual Meeting of the Association of Computational Linguistics*, pages 251–261.

Walker, Marilyn A. 1996. The Effect of Resource Limits and Task Complexity on Collaborative Planning in Dialogue. *Artificial Intelligence Journal*, 85(1–2):181–243.

Walker, Marilyn A. and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proc. 28th Annual Meeting of the ACL*, pages 70–79.

Wang, Michelle and Julia B. Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

Whittaker, Steve and Phil Stenton. 1988. Cues and control in expert client dialogues. In *Proc. 26th Annual Meeting of the ACL, Association of Computational Linguistics*, pages 123–130.

Whittaker, Steve and Phil Stenton. 1989. User studies and the design of natural language systems. In *Proc. 4th Conference of the European Chapter of the ACL, Association of Computational Linguistics*, pages 116–123.

Young, Michael R. 1997. *Generating concise descriptions of complex activities*. Ph.D. thesis, University of Pittsburgh.